ICASE

STABILITY OF GAUSSIAN ELIMINATION

WITHOUT PIVOTING ON TRIDIAGONAL TOEPLITZ MATRICES

Max D. Gunzburger
and
R. A. Nicolaides

Report No. 81-39
December 2, 1981

INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE AND ENGINEERING
NASA Langley Research Center, Hampton, Virginia 23665

Operated by the

UNIVERSITIES SPACE USRA RESEARCH ASSOCIATION

STABILITY OF GAUSSIAN ELIMINATION WITHOUT PIVOTING ON TRIDIAGONAL TOEPLITZ MATRICES

Max D. Gunzburger
Department of Mathematics
University of Tennessee, Knoxville

and

R. A. Nicolaides
Department of Mathematics
University of Connecticut

ABSTRACT

Using the simple vehicle of tridiagonal Toeplitz matrices, the question of whether one must pivot during the Gauss elimination procedure is examined. An exact expression for the multipliers encountered during the elimination process is given. It is then shown that for a prototype Helmholtz problem, one cannot guarantee that elimination without pivoting is stable.

This research was supported by the Army Research Office under Contract No. DAGG-80-K-0056, the Air Force Office of Scientific Research under Grant No. AFSOR-80-91 and by NASA Contract No. NAS1-15810 while the authors were in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA 23665.

1. MULTIPLIERS IN GAUSS ELIMINATION

It has been conjectured that when Gauss elimination is applied to the linear algebraic systems resulting from discretizations of Helmholtz type differential equations,[1] the elimination process may proceed without the need for pivoting for small enough grid sizes. The main goal of this note is to show that in general this conjecture is false. This is not a question of the near singularity of the matrix in the case of the frequency parameter ω being near an eigenvalue of the discrete Laplacian operator. The results below apply most particularly to the case where ω is well away from such an eigenvalue so that the matrix in question is not even approximately singular. The vehicle we use is that of tridiagonal Toeplitz matrices, and the need for pivoting is studied by examining the multipliers encountered when the elimination process proceeds without pivoting.

Consider the Toeplitz tridiagonal matrix

$$\begin{bmatrix} a & b & & & \\ c & a & b & & \\ & \ddots & \ddots & \ddots & \\ & & c & a & b \\ & & & c & a \end{bmatrix} \tag{1}$$

When $a \neq 0$ this matrix is singular only if $4bc > a^2$ and $\cos[j\pi/(n+1)] = -a/2(bc)^{1/2}$ for some integer j between 1 and n. If $a = 0$, the matrix is also singular whenever n is odd. It is easily shown, e.g. by induction, that if we attempt to reduce this matrix to upper bidiagonal form without any pivoting, then the multiplier m_j encountered when we use the j*th* row to eliminate the $(j+1,j)$ element of (1) satisfies the difference equation

$$m_j = c/(a - m_{j-1}b), \quad j = 1,2,\ldots,n-1; \quad m_0 = 0 \tag{2}$$

where n is the dimension of the matrix (1). Further, it can be shown, e.g. again by induction, that these multipliers may be expressed in the form

$$m_j = cE_{j-1}/E_j \ , \ j = 1, 2, 3,\ldots, n - 1 \tag{3}$$

where the E_j's satisfy the linear recurrence relation

$$E_j - a\,E_{j-1} + bc\,E_{j-2} = 0 \ , \ j = 2, 3,\ldots, n - 1 \ ; \qquad E_0 = 1, \ E_1 = a \ . \tag{4}$$

Substituting $E_j = \lambda^j$ in (4) yields that

$$\lambda_{1,2} = \frac{1}{2}[a \pm (a^2 - 4bc)^{1/2}] \ .$$

Then, using the initial conditions $E_0 = 1$ and $E_1 = a$ yields

$$E_j = \frac{\lambda_1^{j+1} - \lambda_2^{j+1}}{\lambda_1 - \lambda_2} \ , \quad j = 0, 1, 2,\ldots \ .$$

Then, from (3)

$$m_j = c(\lambda_1^j - \lambda_2^j)/(\lambda_1^{j+1} - \lambda_2^{j+1}) \ , \ j = 1, 2, 3,\ldots, n - 1 \ .$$

Letting $\lambda_1 = \exp(\alpha + \beta)$ and $\lambda_2 = \exp(-\alpha + \beta)$ we are easily led to

$$m_j = (c/b)^{1/2} \sinh[\alpha j]/\sinh[\alpha(j+1)], \qquad (5)$$

$$j = 1, 2, 3, \ldots, n-1 ,$$

where

$$\cosh \alpha = a/2(bc)^{1/2} . \qquad (6)$$

Formulas (5) and (6) are valid for general complex a, b, and c.

For a, b, c real there are three cases to consider. The first case is when $bc > 0$ and $4bc > a^2$. In this case α is imaginary and (5), (6) become

$$m_j = (c/b)^{1/2} \sin[\beta j]/\sin[\beta(j+1)] \qquad (7)$$

and

$$\cos \beta = a/2(bc)^{1/2} . \qquad (8)$$

Now the angle β is real. The second case is when $bc > 0$ and $a^2 > 4bc$ for which (5), (6) apply with α real. The third case is that of $bc < 0$ for which $\alpha = \gamma - i\pi/2$ with γ real. Then (5), (6) become

$$|b/c|^{1/2} m_j = \begin{cases} \sinh[\gamma j]/\cosh[\gamma(j+1)] & \text{for } j \text{ even} \\ \cosh[\gamma j]/\sinh[\gamma(j+1)] & \text{for } j \text{ odd} \end{cases} \qquad (9)$$

and

$$\sinh \gamma = a/2|bc|^{1/2} . \tag{10}$$

The border case between the first and second case, i.e. $a^2 = 4bc$, yields that

$$m_j = (c/b)^{1/2}\, j/(j+1) \tag{11}$$

while the border case between the second and third cases, i.e. $bc = 0$, yields that

$$m_j = c/a . \tag{12}$$

Note that if $b = 0$, (1) is lower bidiagonal while if $c = 0$, (1) is upper bidiagonal.

We note that if the elimination proceeds without pivoting, we have that the matrix (1) has the factorization

$$\begin{bmatrix} a & b & & 0 \\ c & a & b & \\ & \ddots & \ddots & \ddots \\ 0 & & c & a \end{bmatrix} = \begin{bmatrix} 1 & & & 0 \\ m_1 & 1 & & \\ & \ddots & \ddots & \\ 0 & & m_{n-1} & 1 \end{bmatrix} \begin{bmatrix} u_1 & b & & 0 \\ & u_2 & b & \\ & & \ddots & \ddots \\ 0 & & & u_n \end{bmatrix}$$

where $u_1 = a$ and $u_j = a - b\, m_{j-1}$ for $j = 2, 3, \ldots, n$. Clearly the stability of the elimination process is controlled by the size of

the multipliers m_j for if the m_j's are large, there will be accuracy lost in the calculation of the u_j's .

2. PROTOTYPE HELMHOLTZ PROBLEM

Consider the prototype Helmholtz equation for the function $u(x)$

$$-u_{xx} - \omega^2 u = f(x) \ , \ 0 < x < d$$
$$u(0) \ , \ u(d) \ \text{given} \tag{13}$$

where ω is the given constant frequency. A simple centered difference approximation yields a linear system with a coefficient matrix of the form (1) where $a = 2 - \omega^2 h^2$, $b = c = -1$ and $h = d/(n + 1)$. Here we have subdivided the interval $[0, d]$ into $(n + 1)$ equal subintervals of length h . We have, for h sufficiently small, that $0 < a = 2 - \omega^2 h^2 < 2$ and that $bc = 1$ and $a^2 < 4bc$ so that (7), (8) apply.

The eigenvalues of the operator (13) are given by $s^2\pi^2/d^2$ for $s = 1, 2, 3, \ldots$. Now let $k\pi/d < \omega < (k + 1)\pi/d$, i.e. ω^2 is between the k-th and $(k + 1)$-st eigenvalue of (13), so that $k\pi/(n + 1) < \omega h < (k + 1)\pi/(n + 1)$. Recall that n is the dimension of our matrix so that we wish to examine the multipliers m_j , $1 \leq j \leq n - 1$. Now as j ranges from 1 to $(n - 1)$, $(j + 1)\beta$ ranges over an interval at least as large as

$$[2(k + 1)\pi/(n + 1) \ , \ (n - 1)\ k\pi/(n + 1)] \ .$$

Thus for n large, i.e. h small, $(j + 1)\beta$ ranges over an interval which includes the first k multiples of π. This remains true no matter how small h becomes. Thus regardless of how small h is, the possibility exists that for some j, $(j + 1)\beta$ may be very close to a multiple of π; indeed it may equal such a number. Therefore some multiplier m_j, given by (7), may become arbitrarily large. Note that if $k = 0$ so that $\omega < \pi d$, i.e. the problem (13) is positive definite, then $\beta(j + 1) < \pi$ for $1 \leq j \leq n - 1$ so that the multipliers cannot become large. Of course, for a given ω and h, the multipliers may be well behaved, even if ω is such that our problem is indefinite. However, as indicated above, in general this cannot be guaranteed.

Further insight may be gained by considering perturbations of the parameter ω. Suppose ω and h are such that $\beta = \pi/(\ell + 1)$ exactly for some ℓ such that $1 \leq \ell \leq n - 1$. For h small we have $\beta \cong \omega h = \omega d/(n + 1)$ and therefore

$$\omega d/(n + 1) \cong \pi/(\ell + 1) \quad \text{or} \quad (\ell + 1) \cong \pi(n + 1)/\omega d .$$

Therefore for h small, n is large and $\ell = O(n) = O(1/h)$. For such β, we have $m_\ell = \infty$. Now let us perturb the frequency ω; we let

$$\omega_1 = \omega(1 + \varepsilon\omega')$$

where $\omega' = O(1)$ and $\varepsilon << 1$. It is then easy to show that

$$m_\ell = \left[\frac{\ell}{\ell+1} - \frac{1}{(\ell+1)\omega'\varepsilon}\right]\left[1 + O(h^2 + \varepsilon^2 + \varepsilon h + \varepsilon^2/h)\right] \quad (14)$$

and

$$m_{\ell=1} = \frac{2 - (\ell-1)\omega'\varepsilon}{1 - \ell\omega'\varepsilon} + O(\varepsilon^2 + h^2 + \varepsilon h + \varepsilon^3/h) \quad (15)$$

Clearly as $\varepsilon \to 0$, $\omega_1 \to \omega$, $m_\ell \to \infty$ and $m_{\ell-1} \to 2$.

We now balance ε against $1/h$. First, choose $\varepsilon = O(h^2)$, that is, let $\omega_1 = \omega(1 + \omega' h^2)$. Recalling that $\ell = O(1/h)$, we have from (14), (15)

$$m_\ell = \ell/(\ell+1) + O(1/h) \quad \text{and} \quad m_{\ell-1} = 2 + O(h)$$

so that $m_{\ell-1}$ is still bounded while m_ℓ, which was infinite before we perturbed ω, is now $O(1/h)$. Perhaps this is tolerable if h is not too small.

Now choose $\varepsilon = O(h)$, i.e. $\omega_1 = \omega(1 + \omega' h)$. Then (14), (15) yield that

$$m_\ell = \frac{\ell}{\ell+1} - \frac{1}{(\ell+1)h\omega'} + O(h), \text{ and } m_{\ell-1} = \frac{2 - (\ell-1)h\omega'}{1 - \ell h\omega'} + O(h).$$

If we choose $\omega' = 1/2d$ we have that $\ell h\omega' < \frac{1}{2}$ so that $m_{\ell-1}$ as well as m_ℓ are bounded independent of h. Thus a choice of $\omega_1 = \omega(1 + h/2d)$ will reduce m_ℓ without causing a catastrophe with $m_{\ell-1}$.

It is easy to show that if $u(x; \omega)$ is the solution of (13), then if we vary ω we have that

$$u(x;\ \omega_1) - u(x;\ \omega) = O(\omega_1 - \omega) \ . \tag{16}$$

Therefore, if we choose $\epsilon = O(h^2)$ we have that

$$u(x;\ \omega_1) - u(x;\ \omega) = O(h^2) \ ,$$

an error which is of the same order as the discretization error of the scheme employed above. Therefore we may reduce an infinite multiplier to one with magnitude of $O(1/h)$ by perturbing the frequency in such a manner so that any error introduced is of the same order as the discretization error. On the other hand, if we wish to reduce an infinite multiplier to one with magnitude of $O(1)$, then we must perturb the frequency, and by (16) the solution, by $O(h)$, an error larger than the discretization error.

3. UPWIND DIFFERENCES

We briefly examine a second example which may also be analysed using the well known von-Neumann stability theory. Consider the prototype convection-diffusion equation[2]

$$\begin{aligned} &\frac{1}{R}\, u_{xx} - V\, u_x = f(x) \quad \text{in} \quad 0 < x < d \\ &u(0),\ u(d) \quad \text{given;} \quad V > 0 \ . \end{aligned} \tag{17}$$

If we approximate both u_{xx} and u_x by central difference quotients

we are led to a system with coefficient matrix of the form (1) with $a = -2/Rh^2$, $b = 1/R\,h^2 - V/2h$ and $c = 1/Rh^2 + V/2h$. Thus $bc = [1 - (RVh/2)^2]/(Rh^2)^2$. If $RVh/2 < 1$, i.e. the well known[2] "cell Reynolds number" condition is satisfied, then $bc > 0$ and $a^2 > 4bc$ so that (5), (6) apply and the multipliers are well behaved. (This conclusion can of course also be reached by noting that if $RVh/2 < 1$, the matrix is diagonally dominant.) On the other hand, if $RVh/2 > 1$ so that the cell Reynolds number condition is violated, $bc < 0$ and the multipliers may become large. This is easily seen in the limit $R \to \infty$ (with V, h fixed) for which $m_1 \to \infty$.

Now consider an "upwind differencing" scheme[2] in which u_x is approximated by the backward difference $(u_j - u_{j-1})/h$. We are then led to $a = -2/Rh^2 - V/h$, $b = 1/Rh^2$ and $c = 1/Rh^2 + V/h$. Then $bc = (1/Rh^2)^2(1 + RVh) > 0$ and $a^2 > 4bc$ for all R, h and V . Therefore (5) and (6) apply and the multipliers are well behaved. In the limit $R \to \infty$ it is easy to show that the multipliers tend to unity. Thus, there is no cell Reynolds number condition when upwind differencing is used on the convection term. We note that once again this conclusion may be deduced from the diagonal dominance of the matrix.

REFERENCES

1. Sommerfeld, A., Partial Differential Equations, Academic Press, 1949, New York.

2. Roache, P., Computational Fluid Dynamics, Hermosa Publ., Albuquerque.